Multivariate Time Series Modeling for Forecasting Sintering Temperature in Rotary Kilns Using DCGNet

Xiaogang Zhang, *Member*, *IEEE*, Yanying, Lei, Hua Chen, Lei Zhang,
Yicong Zhou, *Senior Member*, *IEEE*

***Abstract*—The sintering temperature (ST) is a critical index for condition monitoring and process control of coal-fired equipment and is widely used in the production of cement, aluminium, electricity, steel and chemicals. The accurate prediction of the ST is important for control systems to anticipate tragedies. In this paper, we propose a deep learning model for forecasting the ST using automatic spatiotemporal feature extraction from multivariate thermal time series. A hybrid deep neural network named deep convolutional neural network and gated recurrent unit network (DCGNet) is designed to extract multivariate coupling and nonlinear dynamic characteristics for forecasting the ST. DCGNet uses convolutional neural networks (CNNs) and gated recurrent unit (GRU) to extract the local spatial-temporal dependency patterns among the multivariates, and another parallel GRU using historical ST data as input is incorporated to more accurately capture the dynamic characteristics of ST time series. Based on real-world data, application results show that the proposed approach has high forecasting accuracy and robustness, thus having broad application prospects in industrial processes.**

***Index Terms*— Temperature forecasting; Multivariate time series; Convolutional neural network; Gated recurrent unit network.**

I. Introduction

In coal-fired facilities, such as rotary kilns, boilers, and oxygen furnaces, monitoring combustion processes and taking appropriate steps to keep the kiln in stable production conditions are vital to enhancing productivity and reducing exhaust gas and particle emissions.

With the development of information technologies, the measurement and detection of combustion in rotary kilns using soft measurement technologies has frequently been carried out. Because of the complex physical and chemical reactions, heat and mass transfer and multiphase fluid flow occurring simultaneously during sintering, mechanism modeling[1, 2] of sintering is hard to construct. Compared to mechanism modeling, data-driven modeling has exhibited great potential for describing the complex behaviour in the sintering process. By extracting the nonlinear characteristics of images from a charge-coupled device (CCD) camera[3] and thermal process data from a distributed control system (DCS)[4], many data-driven models have been proposed to address online monitoring and prediction for rotary kilns, such as burning condition recognition[5], coal feeding state prediction[6], cement fineness estimation[7], clinker free lime content estimation[8], process fault detection[9], and decomposition rate evaluation of cement raw meal[10]. These models include multilayer perceptrons[7], random vector functional link (RVFL) networks[8], locally linear neuro-fuzzy (LLNF) network[9], generalized regression neural network (GRNN)[4], least squares support vector machines (LS-SVM)[10], radial basis function (RBF) neural network[4, 11], and kernel extreme learning machines (KELM)[12].

The modeling methods mentioned above have made significant achievements in monitoring and predicting combustion, but there is still room for improvement. The combustion process of rotary kilns is a complex nonlinear dynamic system, and the thermal data collected from process sensors are multivariate time series with typical strong coupling and nonlinear dynamic characteristics. Most data-driven models implement static modeling or autoregressive statistical-based prediction considering the information in a separate spatial or temporal scale of variates, and the variates are directly fed into the statistical classifiers or regression without mining their relationships and dynamic dependencies. The precise prediction of the sintering temperature (ST) using thermal data in the field is difficult.

In recent years, deep learning (DL) has made great achievements in its ability to extract hierarchical representations from input data with nonlinear characteristics in various recognition and prediction applications. Convolutional neural network (CNN)[13] have been widely used in image processing and action recognition due to the unique ability of these networks to extract autonomous local shift-invariant characteristics from image data[14]. Furthermore, CNN can

This work was supported in part by the National Key R&D Program of China (No. 2018YFB1305900), and in part by the Natural Science Foundation of Hunan Province, China (No. 2018JJ2056). (Corresponding author: Hua Chen)

X. Zhang and Y. Lei are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: zhangxg@hnu.edu.cn, reylayrey@163.com).

H. Chen is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: chua@hnu.edu.cn).

L. Zhang is with the College of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China (e-mail: lyzl2010@sina.com).

Y. Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

also be used for capturing global information from multidimensional data and noise removal[15].

Recurrent neural network (RNN)[16] have made great progress in time series prediction, such as in natural language processing and speech recognition, because of the ability of these networks to extract temporal information from irregular trends in time series data[17]. Long short-term memory (LSTM) networks[18], as variants of RNN, improve the relatively long-term dependency abilities and solve the training gradient explosion and vanishing problems of RNN[19]. To decrease model complexity and better inhibit the overfitting of LSTM, gated recurrent unit (GRU)[20] have been proposed and used in various applications to handle time series data with irregular changes[21].

Some researchers started to extract multivariate spatial and temporal dependencies by incorporating CNN in RNN models. In [22], a method based on spatiotemporal feature fusion of supervisory control and data acquisition by CNN and GRU was proposed for the condition monitoring of wind turbines. Using CNN to extract deep features from a sequence of frames and inputting these features into bi-directional LSTM network to learn an informative and a non-informative sequence of frames in the cloud-based tier was proposed in [23]. [24] proposed a CNN-LSTM neural network to extract spatial and temporal features to effectively predict housing energy consumption.

In this paper, taking advantage of recent developments in DL research, we propose a novel framework for forecasting the ST in a rotary kiln. We propose a hybrid deep network model named deep convolutional neural network and gated recurrent unit networks (DCGNet) to extract multivariate coupling and nonlinear dynamic characteristics and forecast the ST after the optimal correlative thermal data are selected by using a principal component analysis (PCA) algorithm. A module combines CNNs with GRU network to discover local spatiotemporal dependencies of multidimensional variates, and another GRU network using historical ST data as input is incorporated in parallel to other module to capture the nonlinear dynamic characteristics of the ST. Finally, the deep features extracted by the two modules are weighted fused to feed into a fully connected (FC) layer for forecasting the ST.

The contributions of this paper are as follows:

1) A data-driven model for ST prediction based on DL is proposed. To the best of our knowledge, our work is the first to forecast the ST in rotary kilns by using multivariate thermal process signals. Unlike traditional data-driven modeling methods for recognizing and predicting combustion, a deep hybrid dynamic network model combined with correlated feature selection is utilized for predicting the ST. This DL-based framework provides a new idea for establishing a soft sensor model of industrial process data with large time lags, multivariate coupling and nonlinear characteristics.
2) Unlike conventional static state-based and autoregressive statistics-based methods, a hybrid deep network, DCGNet, with a parallel concatenated structure based on CNNs and GRU networks is proposed to extract the coupling and nonlinear dynamic characteristics of multivariate thermal process data. The concatenation of CNNs and GRU network in the first module is used to extract the local nonlinear coupling and dynamic characteristics from multiple process data, and another parallel GRU network is adopted to capture the long-term dynamic dependency of historical ST data. The FC layer connects the two parallel modules for single-step forecasting of the ST.

This paper is organized as follows: Section II describes the technological process of rotary kilns and formulates the problem. Section III describes PCA and the structure of the proposed DCGNet in detail. Section IV illustrates the experimental results. Finally, the conclusion is drawn in Section V.

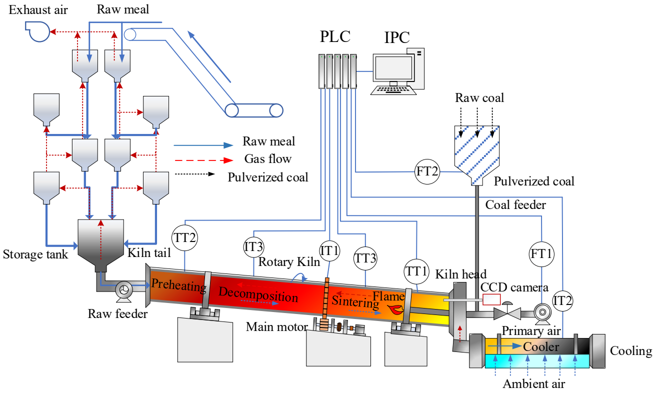

Fig. 1. Rotary kiln production process

II. Process Description and Problem Formulation

The process flow chart of an alumina rotary kiln is shown in Fig. 1. The rotary kiln barrel is 3.5-4.5 m in diameter and 90-110 m in length and has an average daily output of several tons. During the production process, the kiln is installed obliquely and driven by a motor rotating slowly. High-silica bauxite, soda ash, lime, etc., are mixed and ground into a raw material slurry in a certain proportion, and then the slurry flows from the kiln tail to the kiln head. Additionally, the coal powder is blown into the kiln by a blower from the kiln head to the kiln tail end. The coal powder and the material burn in the sintering zone in the high-temperature environment, which is as high as 1100-1300 °C under normal sintering conditions. The material is dried, preheated, and sintered before entering the cooling machine. The process of burning material in the sintering zone (burning zone) is called sintering.

The sintering zone is approximately 10-15 m away from the kiln head. The ST in the sintering zone is an important index for combustion monitoring and control because the ST can reflect the combustion conditions to a certain extent, while conventional high-temperature physical measurement sensors, such as thermocouples and ultraviolet sensors, are difficult to deploy due to either the rotational structure of the rotary kiln or the dust and particles inside the kiln. An infrared thermometer can measure the ST by a temperature-sensing element installed in the burning zone. However, the temperature-sensing element is prone to inaccurate measurement due to optical pollution or optical deviation[25]. An infrared camera system via

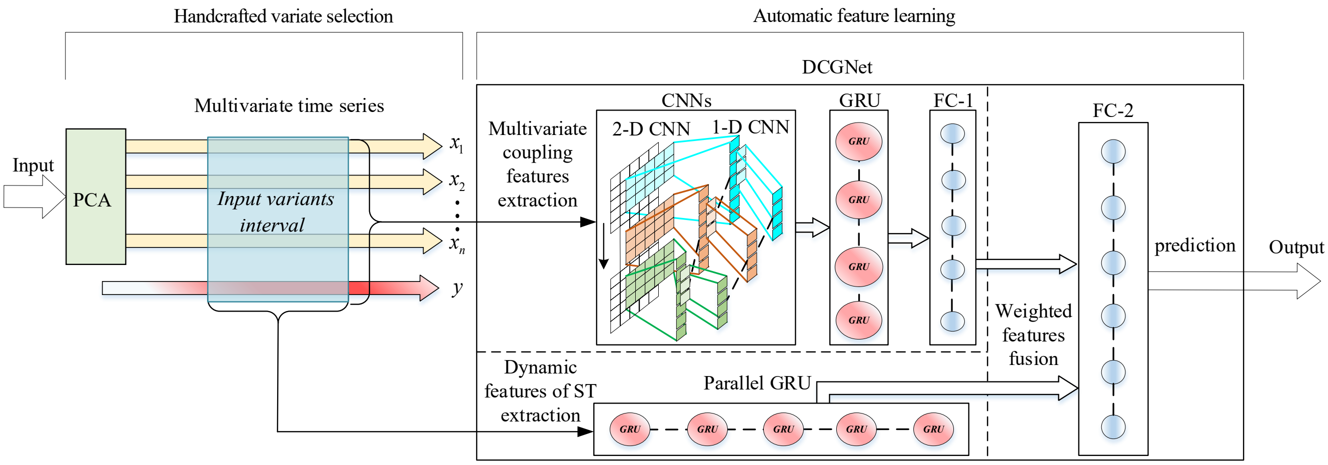

Fig. 3. The algorithm framework based on the DCGNet model

vision-based algorithms can be used for ST measurement owing to the intuition and immediacy of the infrared camera system, while the performance of the system is influenced by factors such as emissivity, scattering and absorption of the light path, and background noise in actual temperature measurement[12].

The sintering process in a rotary kiln has been considered one of the most complicated processes because complex physical and chemical reactions, heat and mass transfer and multiphase fluid flow occur simultaneously. For example, in the sintering of alumina, more than fifty material substances are used in approximately twenty types of physical and chemical reactions in the kiln. Three main characteristics of sintering in a rotary kiln are as follows:

1) Dynamic nonlinearity: Because of the complex physical and chemical reactions, the relationship among thermal process variates is complicated and difficult to clearly express using mathematical models. The change in a variate is influenced by changes in other variates.
2) Multivariate coupling: During sintering in a rotary kiln, there are some observed variates, such as the kiln head temperature, kiln tail temperature, main driven current, and ST. According to changes in these observed variates, the operators and control system will adjust the operational variates, such as the coal feeding rate and blast flow rate, to maintain the kiln under normal sintering conditions. In turn, the adjustment affects the observed variates. Therefore, these process variates interact with each other, and there are correlations between their values.
3) Large time lag: Because of the large size and the slow heat transfer mechanism of the kiln, the ST is influenced by previous thermal process variants over a certain period. After the state variates (such as the kiln head temperature) change, the ST will not change immediately. The change in the ST is the cumulative response of all variables over time. Forecasts of the ST should consider the time lag and accumulative effect between the previous thermal process variants and the ST.

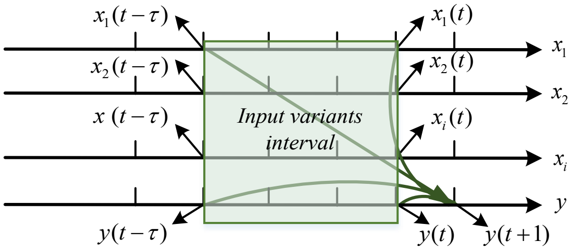

Fig. 2. Illustration of the model for predicting the ST

In this paper, time series forecasting is focused on multiple input variates and a singular prediction output. More formally, a series of process variates $X=\{x_1,\ldots,x_i\}$ and y as the variate to be estimated (ST) are given, where $X\in\mathbb{R}^{N\times i}$, $y\in\mathbb{R}^{N}$, i is the variable dimension and N is the number of samples collected by the DCS. We aim to predict the next signals in a rolling forecasting fashion. Assuming that $\{X(t),\ldots,X(t-\tau),$ $y(t),\ldots,y(t-\tau)\}$ are available, we estimate the ST at the next moment, thus predicting $y(t{+}1)$, where τ is the correlative time interval of the thermal process data. The nonlinear dynamic mapping is formulated as follows:

$$y(t+1)=f(X(t),\ldots,X(t-\tau),y(t),\ldots,y(t-\tau)) \tag{1}$$

As shown in Fig. 2, the prediction of the ST is implemented using previous thermal process data in a period with a fixed length of τ. The initial model is constructed as a nonlinear function from the multidimensional thermal process variate X and ST y.

III. Modeling Methodology

Focusing on the abovementioned practical modeling problem, a novel DCGNet-based modeling method is proposed in this paper for predicting the ST in rotary kilns, as shown in Fig. 3. First, to reduce the model complexity and improve the prediction performance, PCA is introduced to select the optimal model inputs. Then, unlike conventional nonlinear dynamic process system modeling, which adopts an overall model containing all the variates, the proposed DCGNet separately models the selected variates and historical ST data. In the first module, CNNs combined with GRU network are used to capture coupling features and local temporal information of multivariates. In the second module, a parallel GRU network models historical ST information and captures the nonlinear dynamic characteristics of the ST time series. Finally, an FC layer is used to fuse these two modules and predict the ST. Through the construction of a three-part model, more abundant characteristic information can be obtained, and the modeling accuracy of the constructed DCGNet model can be further enhanced.

A. Variate selection based on PCA

In industrial applications, due to the redundancy between input variables, if all input variables are used as model inputs, the complexity will increase, and the performance of the model will decrease. When the number of learning samples is large, PCA [26] can be utilized to select a few key correlated factors from all process variates as the inputs of ST modeling.

PCA is an effective feature extraction method that replaces the original variates with low-dimensional principal components, thereby filtering out noise and reducing the dimensionality of the variates, while the nonlinear principal component obtained by PCA is essentially the combination function of the original variates. These low-dimensional principal components used as the input of the model have no clear physical meaning and no theoretical interpretability; therefore, these components are unsuitable for actual sintering process control of a rotary kiln. Nevertheless, the component matrix obtained by PCA indicates the correlation between the original variates and each principal component, and the optimal input variates can be selected according to the components.

B. Framework of DCGNet

After the original number of thermal process variates is reduced by PCA, the selected input variate data between time $t-\tau$ and time t are input into DCGNet. The concatenated CNNs and GRU are designed to extract the multivariate coupling and nonlinear dynamic characteristics. Additionally, to more precisely discover the nonlinear dynamic characteristics of the ST, another parallel GRU is used solely for the historical ST. By combining these two modules, precise deep features can be extracted automatically as the inputs of the FC layer; then, the ST at time $t+1$ is forecast by the FC layer as output. In the following sections, we introduce each network layer of DCGNet in detail in three parts.

1) Multivariate coupling feature extraction module

Focusing on the coupling characteristics of the multiple process variates, two CNN layers are utilized. The selected thermal data are first input into a two-dimensional convolutional layer (2-D CNN) without pooling, thereby aiming to extract local spatial coupling correlations among the multivariates by the two most important features of the CNN: local perception and weight sharing.

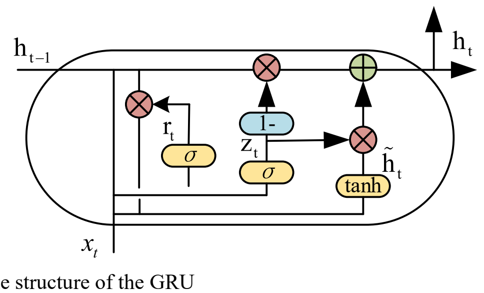

Fig. 4. The structure of the GRU

The inputs of the CNN layer are a two-dimensional matrix $Xy=[[X(t),...,X(t-\tau)],[y(t),...y(t-\tau)]]$ with a size of $(\tau+1)\times(n+1)$. n is the number of optimal process variates selected by PCA. The 2-D CNN consists of multiple filters whose width is set to be the same as the number of variates. The filters sweep through the input matrix Xy and produce

$$h_{il}^{N}=\mathrm{ReLU}(w_{j\times(n+1)}^{N}*Xy+\beta^{N}) \tag{2}$$

where * denotes the convolution operation and $w_{j\times(n+1)}^{N}$ is the weight of the *N*-th filter of size $j\times(n+1)$. β^{N} is the bias of the convolution layer. The ReLU function is $\mathrm{ReLU(x)}=\max(0,\mathrm{x})$. In the case of ignoring the boundary filling, the output of the 2-D CNN has a size of $1\times l$, where $l=\frac{(\tau+1)-j}{s}+1$ and s is the step length.

To further refine the features while reducing the number of model parameters and improving the operation speed, the multiple filters as a one-dimensional convolutional layer (1-D CNN) sweep through from top to bottom in a certain step on the vectors again, and then the space-time coupling features h_{il} are compressed as follows:

$$H^{K}=\mathrm{ReLU}(\sum_{i=1}^{N}w^{K}*h_{il}^{N}+\beta^{K}) \tag{3}$$

where K denotes the number of filters. After two convolution calculations, more refined spatial-temporal coupling features of multiple process variates are produced.

The GRU (Fig. 4) is designed with memory units and gate functions in the perceptron to memorize the historical information, thereby overcoming the problems of capturing long-term dependencies encountered by the RNN.

After the feature extraction of two convolutional layers, a GRU layer is added to store time information about important characteristics of multivariate variates. The last hidden state of the GRU cell at time t is computed as:

$$r_{t}=\sigma(w_{r}h_{t-1}^{Xy}+v_{r}H_{t}+\beta_{r}) \tag{4}$$

$$z_{t}=\sigma(w_{z}h_{t-1}^{Xy}+v_{z}H_{t}+\beta_{z}) \tag{5}$$

$$\tilde{h}_{t}=tanh(w_{h}(r_{t}\odot h_{t-1}^{Xy})+v_{h}H_{t}+\beta_{h}) \tag{6}$$

$$h_{t}^{Xy}=((1-z_{t})\odot h_{t-1}^{Xy}+z_{t}\odot\tilde{h}_{t}) \tag{7}$$

where w and v are weight matrices, β is the bias vector, $tanh$ is a hyperbolic tangent function, $\odot$ is an element-wise multiplication, and σ is a sigmoid function.

Eqs. (4) and (5) show the operation of the reset gate r_t and the update gate z_t. After the output at the current moment H_t and the hidden state of the previous time step h_{t-1}^{Xy} are obtained, the probability of updating or resetting is determined by a sigmoid function. In Eq. (6), the current data at the same time step H_t and the partial past hidden state selected by the reset gate r_t determine the new memory content $\tilde{h}_t$ by nonlinear change. In Eq. (7), the update gate z_t filters the new memory content $\tilde{h}_t$ and the hidden state at the previous time step h_{t-1}^{Xy} to form the current dynamic coupling information h_t^{Xy}.

After the GRU layer, FC-layer-1 unearths the nonlinear representation of the spatiotemporal coupling features and obtains more precise deep feature information. The output of the series GRU layer H^{Xy} is computed as:

$$D^{Xy} = f(w_{Xy}H^{Xy} + \beta_{Xy}) \tag{8}$$

where w_{Xy} and β_{Xy} are the weight matrix and bias vector, respectively, in FC-layer-1. D^{Xy} is the nonlinear space-time coupling feature of multivariate variates.

2) *Dynamic features of the ST extraction module*

A parallel GRU layer, which has the same function as a serial GRU layer but has different inputs, is used solely to capture the nonlinear dynamic characteristics in the ST time series. The historical ST data over a certain period $\{y(t),...,y(t-\tau)\}$ are used as the inputs of the GRU layer. Then, the GRU stores the time information of irregular changes in the early stage of the ST sequential data in the cell through the feedback structure. With each step update, nonlinear dynamic features H^y of the ST are obtained through the function of the parallel GRU layer.

Algorithm 1. Outline of DCGNet training for predicting the ST in a rotary kiln

```
Input: The training set $Z = \{(X, y) \mid X \in \mathbb{R}^{L\times n}, y \in \mathbb{R}^{L}\}$
Output: Weight matrix $w$, weight matrix $v$ and bias vector $\beta$.
Initialization: Determine the deviation threshold $\varepsilon$, and set the hyperparameter; the iteration number $I=0$, and the maximum number of iterations is $I_{\mathrm{itera}}$. Randomly initialize the weight matrix and bias vector.
Repeat:
    Forward Propagation:
    DO
        Step 1. : Conduct a 2-D convolution operation with the multivariate process variable data in Eq. (2). Use GRU Eqs. (4)-(7) to extract nonlinear dynamic features from the ST time series data.
        Step 2. : Conduct a 1-D convolution operation with the coupling features from the 2-D CNN layers in Eq. (3).
        Step 3. : Use GRU Eqs. (4)-(7) to further process the temporal information by using the features extracted from the 1-D CNN layers.
        Step 4. : Use FC-layer-1 to reveal the nonlinear representation of the spatiotemporal coupling features in Eq. (8).
        Step 5. : Use FC-layer-2 to determine the output in Eq. (9).
        Step 6. : Calculate the $Loss$ introduced in Eq. (10) between the prediction and targets.
    end;
    Backward Propagation:
        Compute the gradient by using Adam, and update the weight matrix and bias vector. Let $I=I+1$.
until: If $Loss < \varepsilon$ or $I > I_{\mathrm{itera}}$
```

3) *Weighted feature fusion and prediction*

After extracting features, we first use FC-layer-2 for weighted fusion of the output of the two modules, and the final outputs of DCGNet are then obtained by line nonlinear regression as

$$\tilde{y} = f(w_D D^{Xy} + w_y H^y + \beta_{Fc}) \tag{9}$$

where $\tilde{y}$ denotes the model's final prediction; w_D, w_y and β_{Fc} denote the weight matrices and bias value, respectively; and f is a nonlinear activation function.

C. *Objective function and optimization strategy*

In this paper, the mean squared error (MSE) loss function is used for optimization in model training and is defined as:

$$Loss = \frac{1}{L}\sum_{i=0}^{L}(y-\tilde{y})^2 \tag{10}$$

Where L is the total number of training data in the time sequence and y is the real value.

According to the above framework, the Adam optimization algorithm is used to find the gradient of the network error for each weight parameter in back-propagation, and the new weight is obtained through the parameter update process. The model weights are calculated iteratively until a predetermined small loss is reached, and the optimal predicted value is obtained. The reasons for choosing Adam as an optimizer are that it can design independent adaptive learning rates for different parameters and, most importantly, using Adam makes our calculations more efficient. The test set is substituted into the trained model to predict the ST in the rotary kiln once the training work is performed by using the training set. The entire DCGNet training process is summarized in Algorithm 1.

IV. Experiments and Discussion

A. *Experimental data*

To verify the effectiveness of DCGNet, we collected many thermal data from the on-site thermal instruments of a No. 2 rotary kiln manufactured by the Zhongzhou Aluminium Company in China. A total of 7000 samples, with a sampling interval of 1 min, were collected for prediction and evaluation. According to expert knowledge and the on-site DCS, 7 auxiliary variables closely related to the ST were collected. The detailed descriptions of the variates are listed in

TABLE I
THE THERMAL VARIABLES OF THE ROTARY KILN

Variables	Input variable description	Unit	Mean value
x_1	Kiln head temperature	°C	551.82
x_2	Kiln tail temperature	°C	243.28
x_3	Main motor current	A	259.42
x_4	Cooling fan current	A	226.30
x_5	Air flow	m^3/h	18210.13
x_6	Rotation speed	Rad/min	0.91
x_7	Coal feeding value	Rad/min	8.54
y	Flame temperature	°C	1063.89

TABLE II
ACCUMULATED CONTRIBUTION RATES AND EIGENVALUES CALCULATED BY PCA

Principal components	Accumulated contribution rates (%)	Eigenvalues
1	41.82	0.067
2	68.81	0.043
3	80.44	0.019
4	**91.07**	0.017
5	95.87	0.008
6	98.60	0.004
7	100	0.002

TABLE III
COMPONENT MATRIX

Variables	Principal component			
	1	2	3	4
x_1	0.17938	-0.21316	0.31056	**0.83398**
x_2	-0.01388	-0.23789	-0.11606	-0.16802
x_3	0.00964	-0.09695	0.04392	-0.06284
x_4	0.06084	0.20393	0.22306	0.29287
x_5	0.68296	-0.31904	**0.51467**	-0.38970
x_6	**0.69430**	0.47616	-0.50886	0.11336
x_7	-0.12394	**0.72005**	0.56087	-0.14766

Table I. The ST data were detected by an infrared thermometer. All samples were divided into 75%, 5% and 20% proportions for training, validation and testing, respectively. For standardization, all the data were scaled into the range of 0 to 1, according to Eq. (11), before the model was established, and then the predicted results were transferred back to the same units according to Eq. (12):

$$a' = \frac{a - a_{\min}}{a + a_{\max}} \tag{11}$$

$$a = a'(a + a_{\max}) + a_{\min} \tag{12}$$

where $a_{\min}$ and $a_{\max}$ are the minimum and maximum values, respectively, of the variable and a' and a are the variable parameters after and before scaling, respectively.

B. The optimal input variates and interval parameter selection

The accumulated cumulative contribution rates of the seven variates are shown in Table II; as this table shows, the accumulated cumulative contribution rate of the first four principal components is 91.07%. Further, Table III shows that the variates with the largest weight coefficients among the first four principal components are selected as the best features in the component matrix. That is, the kiln head temperature, air flow, rotation speed and coal feeding value are the optimal input variates.

The interval parameters of the four input variates are determined by a grid search. $\tau = 24$ is selected as the optimal time interval of the model.

C. Forecasting accuracy and performance comparison with other methods

To compare the performance of the DCGNet model, we conducted extensive experiments in which seven methods were used on industrial datasets for ST forecasting. The methods used in our comparative evaluation are as follows:

1) DAE[27]: A denoising autoencoder network with three hidden layers
2) MLP[28]: A neural network with three hidden layers
3) DCNN[29]: A convolutional neural network consisting of one-dimensional and two-dimensional convolutional layers
4) CNN-LSTM[24]: A two-dimensional convolutional neural network combined with a long short-term memory network
5) DLSTM[18]: A two-layer long short-term memory network with dropout
6) DGRU[20]: A two-layer gated recurrent unit network with dropout
7) DBiGRU[30]: A two-layer bi-directional gated recurrent unit network with dropout

The experiments were carried out on a personal computer with an Intel i7-7700 CPU (3.20 GHz) and 8.0 GB RAM. DCGNet and the other neural network algorithms used are based on Pytorch and Python version 2.7.3.

In this paper, we performed cross-validation and a grid search to determine the optimal parameters for each model. All methods use a back-propagation algorithm to continuously adjust the weight matrix and bias between the hidden and output layers. The hyperparameters were adjusted to achieve high performance for each model and compared with those of current methods and DCGNet. We experimentally found that using the Adam algorithm with a learning rate of 0.01 is more conducive to error convergence. With the same input data, each model was trained 20 times, and the average value of the results was used as the experimental result.

To train DCGNet, it is important to determine the key

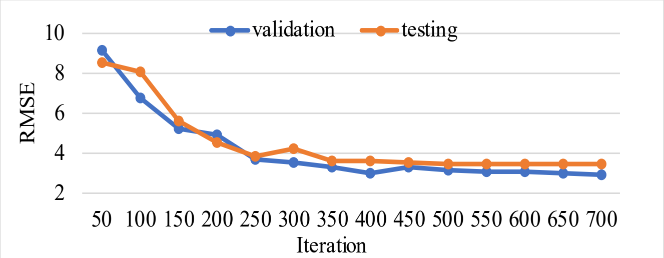

Fig. 5. Comparison of the RMSE at different numbers of iterations for DCGNet

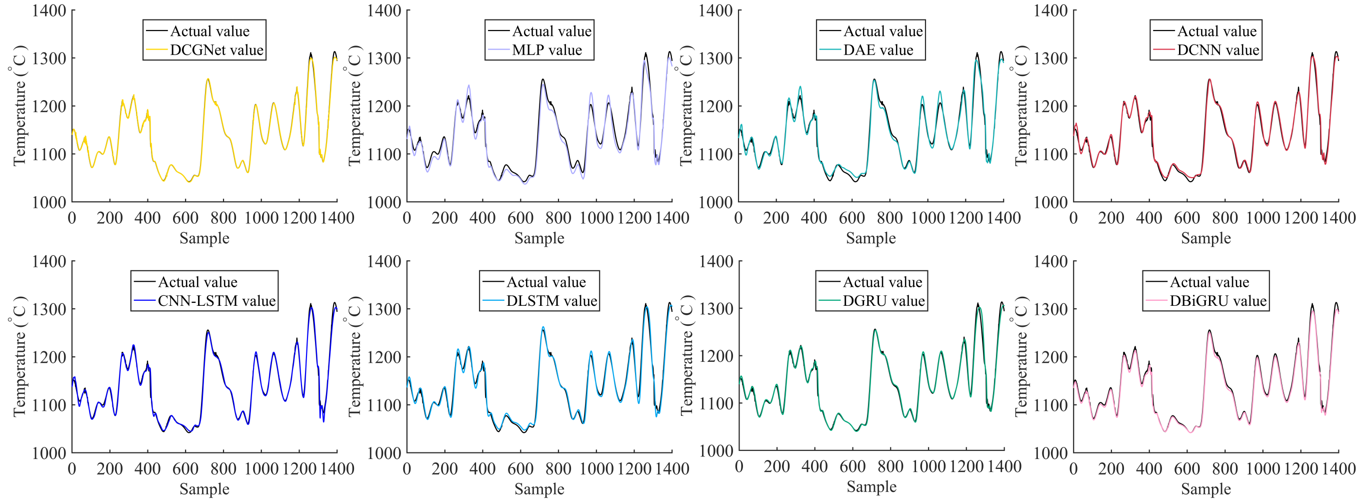

Fig. 6. Prediction of ST by different algorithms

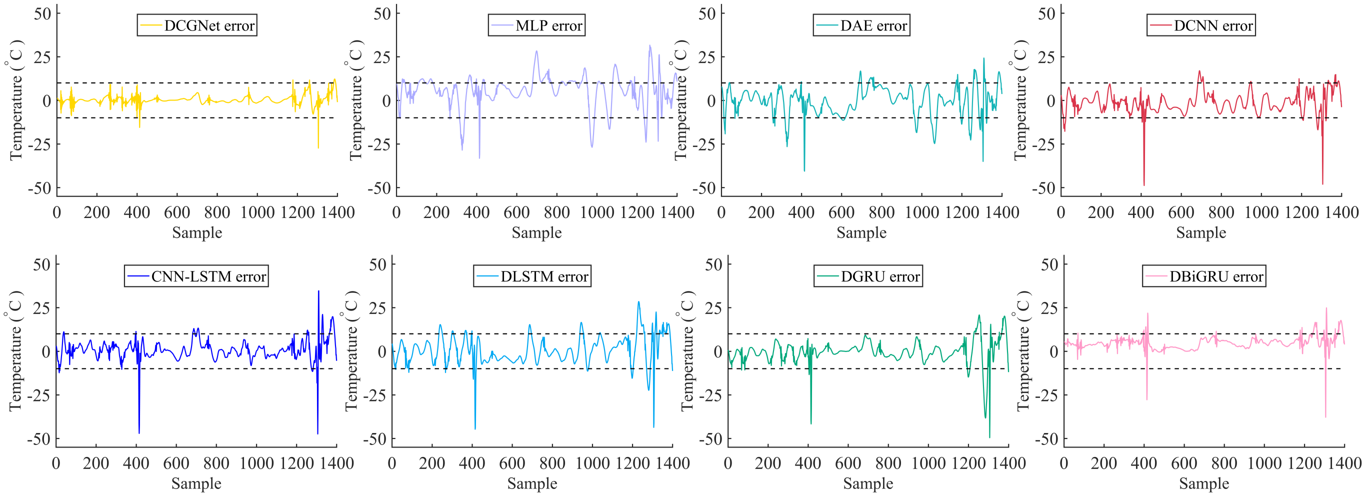

Fig. 7. Prediction error curves of different algorithms

hyperparameter for the number of iterations in back-propagation through time (BPTT). Here, for the validation and test datasets, we researched the prediction performance of the root mean square error (RMSE) under different iteration conditions. As Fig. 5 shows, as the number of iterations increases, the RMSE gradually decreases. When the number of iterations reaches 500, the RMSE reaches a convergence state. Therefore, the number of iterations selected in this study is 500.

For the DAE, the hidden units are chosen as [50,20,50]. For the MLP, the hidden units are chosen as [50,20,50]. For the DCNN, we adopted a ReLU as the activation function, and the chosen hidden unit was [50,50]. Their filter sizes were set to 6×5 and 2×1. For the CNN-LSTM, we chose 80 hidden units for the CNN layer. The filter sizes of this layer were set to 2×1. The pooling layer has the same filter sizes. The LSTM layer size in the CNN-LSTM was 40. For the DLSTM, DGRU, and DBiGRU, we performed dropout after each layer, and the rate was usually set to 0.2. The number of hidden units was set as 50. For DCGNet, we performed dropout after the GRU layer, and the rate was set to 0.2. We have 50 hidden layer units for the two GRU layers. We adopted a ReLU as the activation function, and the chosen hidden unit was [50,50] for the two CNN layers. Their filter sizes were set to 6×5 and 2×1. For the two FC layers, we chose the hidden units as [85,50]. We empirically found that the *tanh* function leads to more reliable performance.

In this paper, the mean absolute error (MAE), RMSE and correlation coefficient (CC) are calculated to estimate the prediction accuracy. The definitions of these metrics are as follows:

$$MAE=\frac{1}{N}\sum_{i=0}^{N}|y-\tilde{y}| \tag{13}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=0}^{N}(y-\tilde{y})^2} \tag{14}$$

$$CC=\frac{Cov(y,\tilde{y})}{\sqrt{Var[y]Var[\tilde{y}]}} \tag{15}$$

where y is the actual value, $\tilde{y}$ is the predicted value, and N is the number of samples. The MAE and RMSE are used to determine the accuracy of the algorithms. The CC represents the correlation between the input and output. For the MAE and RMSE, a lower value is better, while for the CC, a higher value is better.

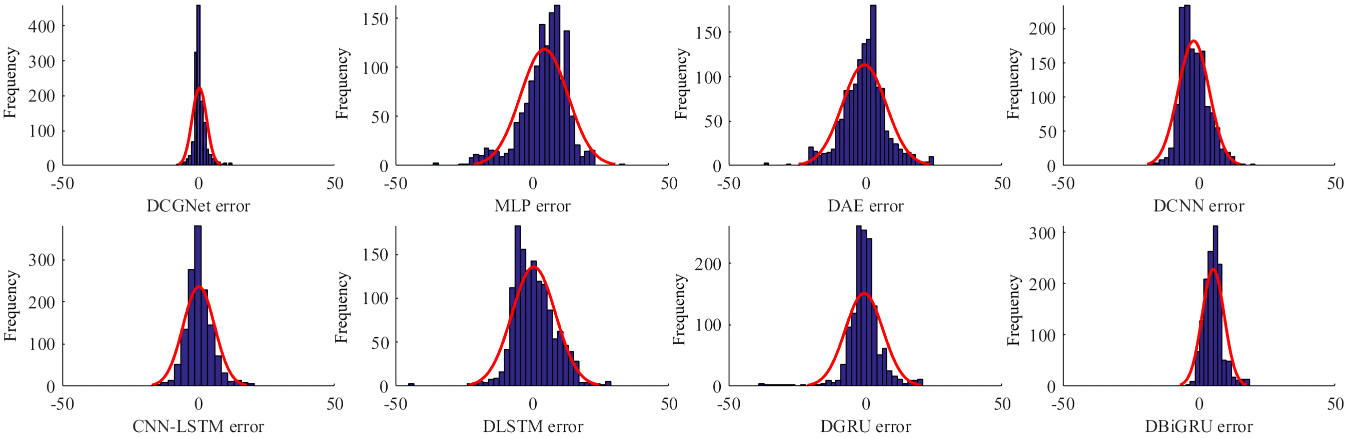

Fig. 8. Error probability distribution curves of different algorithms

Table IV summarizes the evaluation results of all the above-mentioned methods. The actual and predicted values of the eight prediction methods are compared in Fig. 6, and the corresponding prediction error curves are shown in Fig. 7.

Clearly, the prediction accuracy of the traditional MLP and DAE models is the lowest, and the RMSE values of these models are significantly higher than those of the other models. Similar results are evident in Fig. 6. Both models perform poorly, and they cannot easily fit to the trend of local changes in real values.

In comparison with that of the above two models, the prediction accuracy of the three kinds of models containing a dynamic recurrent neural network is significantly improved. As Table IV shows, the RMSE values of the DGRU and DBiGRU models are 5.9534 and 5.8681, respectively. In contrast, the DLSTM-based dynamic networks have poor prediction performance, with an RMSE value of 6.8883. In general, the curves of these three models basically agree with the actual trend.

TABLE IV
COMPARISON OF EXPERIMENTAL RESULTS FOR INDUSTRIAL DATA

Model properties	Methods	MAE	RMSE	CC
Hybrid model	**DCGNet**	**2.0213**	**3.4359**	**0.99881**
	CNN-LSTM	3.8745	5.8152	0.99649
Convolutional network	DCNN	4.6112	6.0277	0.99643
Static network	MLP	7.1573	10.0170	0.98890
	DAE	7.4002	8.9617	0.99412
Recurrent network	DLSTM	4.7459	6.8883	0.99538
	DGRU	4.5449	5.9534	0.99645
	DBiGRU	4.2675	5.8681	0.99638

TABLE VI
COMPARISON OF EXPERIMENTAL RESULTS FOR INDUSTRIAL DATA WITH 5% RANDOM GAUSSIAN WHITE NOISE

Model properties	Methods	MAE	RMSE	CC
Hybrid network	**DCGNet**	**2.0154**	**3.6804**	**0.99846**
	CNN-LSTM	4.1411	6.2840	0.99551
Convolutional network	DCNN	4.5671	6.3514	0.99518
Static network	MLP	8.3487	12.4810	0.98443
	DAE	11.3943	13.9245	0.99571
Recurrent network	DLSTM	6.0978	7.5758	0.99561
	DGRU	4.6499	6.3046	0.99555
	DBiGRU	4.8517	6.6652	0.99503

Furthermore, the proposed DCNN method also performs better than the static neural network, with RMSE and MAE as low as 4.6112 and 6.0277, respectively. As Fig. 6 shows, the predicted value coincides well with the actual value with a smaller deviation. Therefore, the strong coupling relationship between multiple variables cannot be ignored in the modeling of the rotary kiln.

Fig. 6 shows that the CNN-LSTM adequately models the irregular trend of the ST. The CNN-LSTM evidently has the smallest error compared to that of the static and recurrent networks compared in Table IV, thus further verifying the effectiveness of the hybrid combination of a convolutional and dynamic recurrent network. However, compared with the prediction performance of the proposed DCGNet model, the prediction performance of this hybrid network can be further improved.

Finally, Table IV shows that the DCGNet model proposed in this paper has the best prediction accuracy, the minimum values of the RMSE and MAE and the maximum value of the CC, indicating a strong correlation between the real and predicted values. As Fig. 7 shows, compared to the other methods, the proposed method minimizes the error in the same intervals, thereby indicating that the prediction performance of the model is optimal.

In addition, to further evaluate the performance of the different prediction models, we introduce an error probability distribution curve in this paper. Fig. 8 shows the probabilistic distributions of prediction residuals based on the eight different prediction models. Compared with other models, the error distribution of the predicted ST obtained by the model developed in this paper is much closer to zero with smaller variations, thus further indicating the reliability of our proposed method.

TABLE V
THE TRAINING TIME OF DIFFERENT MODELS

DCGNet	DAE	MLP	DCNN	CNN-LSTM	DLSTM	DGRU	DBiGRU
90.9 s	48.1 s	32.1 s	26.1 s	65.6 s	49.2 s	48.1 s	73.4 s

Besides the prediction accuracy, the training time of the prediction model is also very important. the prediction accuracy, the training time of the prediction model is also very important. Table V shows that the DCNN has the shortest training time because the convolution operation reduces the model parameters. In contrast, the rest of the network training times are significantly longer. Improving the DCGNet prediction accuracy involves appropriately increasing the complexity of the model. Therefore, the trade-off between training time and prediction accuracy is acceptable for our proposed model.

D. *Robustness of the models*

Furthermore, to verify the robustness of our proposed DCGNet model, we add 5% Gaussian white noise to the training data. As Table VI shows, increasing the noise by 5% reduces the performance of all the models. However, compared to the performance of other models, the performance of DCGNet remains optimal. Moreover, Tables IV and VI show that our proposed DCGNet model has the smallest variation in the RMSE of the predicted values, with a range of 0.2445. The experimental results demonstrate that our proposed DCGNet model has robust performance with each part together and is suitable for predicting the ST in rotary kilns.

E. *Ablation study of DCGNet*

To demonstrate the validity of the method of separately extracting features and then combining them for prediction, a comparative study is conducted. First, we define the different models as follows:

1) DCGNet-WC: A DCGNet model without two CNN layers
2) CGRU-XY: A DCGNet model without the parallel GRU layer
3) GRU-Y: A parallel GRU layer for the historical ST
4) ARMA-Y: An autoregressive moving average model for the historical ST

For the above models, the test results measured using three evaluation metrics are shown in Table VII. Several conclusions from these experimental results are summarized as follows:

1) The lack of CNN layers reduces the prediction performance. Two CNN layers can effectively capture the multivariate coupling characteristics of rotary kiln sintering.
2) Removing the input of historical ST data in CGRU-XY greatly reduces the performance. The dynamic characteristics of historical ST data cannot be ignored.
3) The sintering process has the characteristics of multivariate coupling so that a variate state is affected by other thermal variables. The univariate methods of GRU-Y and ARMA cannot uncover the latent correlation of time series data and may be unsuitable for time series prediction with irregular fluctuations in a rotary kiln industrial process.
4) All the inputs of DCGRU together and each part of the network lead to the optimal performance of our approach, thus verifying that our method can obtain richer feature information through accurate modeling and further improve the prediction accuracy.

TABLE VII
THE PREDICTED RESULTS IN THE ABLATION STUDY

Methods	MAE	RMSE	CC
DCGNet	**2.0213**	**3.4359**	**0.99881**
DCGNet-WC	3.5206	5.4252	0.99668
CGRU-XY	3.9887	5.6581	0.99649
GRU-Y	4.1636	6.2911	0.99551
ARMA-Y	5.0206	6.9082	0.99528

F. *Input selection experiment*

The prediction performance of DCGNet is compared by choosing all variables as inputs or the best variable selected by PCA as input to verify the effectiveness of the PCA method. According to Fig. 9, when all variates are used, the MAE and RMSE of our proposed DCGNet are 2.8605 and 4.6045, respectively. In contrast, after the variates are selected using PCA, the MAE and RMSE are reduced to 2.0213 and 3.4359, respectively. The comparison results show that PCA can effectively eliminate the redundancy between input variates and improve the prediction accuracy of the model.

The interval τ is a parameter that controls the length of a variate that is input into our proposed DCGNet model. Clearly, a small τ does not fully consider the local sequential signal of the input data, so too many spatiotemporal coupling features may be lost. With a large τ, the presence of signal redundancy increases the computational burden. We tested our proposed DCGNet while using different intervals in combination with sintering. As previously reported, the other parameters of our proposed DCGNet remain unchanged. Fig. 10 shows that a very

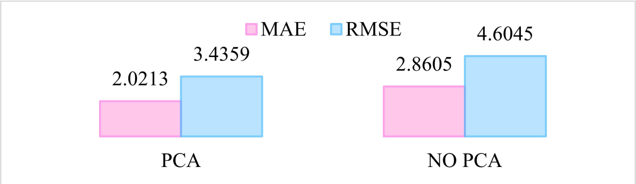

Fig. 9. Experimental comparison of PCA analysis

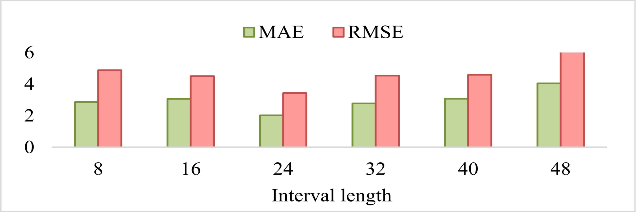

Fig. 10. Experiment with different input time intervals

large and small τ both impede the performance of our proposed DCGNet. With a moderate interval length of τ=24, our proposed DCGNet model achieves the best performance.

V. Conclusion

In this paper, aiming to determine the multivariate coupling and dynamic nonlinear characteristics of rotary kiln sintering, a new data-driven model for ST prediction based on a hybrid deep network is proposed. By using PCA, the optimal correlative variates of the network are selected. Then, the DCGNet model based on CNNs and GRU networks is constructed to learn the deep representations of multivariate thermal process data. Comparative experiments and an ablation study on real-world data verified the effectiveness and robustness of our method.

In future research, we will explore a universal framework for the prediction modeling of other variables of rotary kilns, especially the modeling of control variables, such as coal feeding. Such a framework has great significance for the process optimization of production.

References

[1] X. Liu, X. Xu, W. Wu, F. Herz, and E. Specht, “A simplified model to calculate the power draw for material movement in industrial rotary kilns,” *Powder Technology,* vol. 301, pp. 1294-1298, 2016.

[2] M. S. Manju, and S. Savithri, “Three dimensional CFD simulation of pneumatic coal injection in a direct reduction rotary kiln,” *Fuel,* vol. 102, pp. 54-64, 2012.

[3] W. Li, D. Wang, and T. Chai, “Flame Image-Based Burning State Recognition for Sintering Process of Rotary Kiln Using Heterogeneous Features and Fuzzy Integral,” *IEEE Transactions on Industrial Informatics,* vol. 8, no. 4, pp. 780-790, 2012.

[4] A. K. Pani, and H. K. Mohanta, “Online monitoring of cement clinker quality using multivariate statistics and Takagi-Sugeno fuzzy-inference technique,” *Control Engineering Practice,* vol. 57, pp. 1-17, 2016.

[5] H. Chen, X. Zhang, P. Hong, H. Hu, and X. Yin, “Recognition of the Temperature Condition of a Rotary Kiln Using Dynamic Features of a Series of Blurry Flame Images,” *IEEE Transactions on Industrial Informatics*, pp. 1-1, 2015.

[6] X. Zhang, L. Zhang, H. Chen, and B. Dai, “Prediction of coal feeding during sintering in a rotary kiln based on statistical learning in the phase space,” *ISA Trans,* vol. 83, pp. 248-260, Dec, 2018.

[7] D. Stanisic, N. Jorgovanovic, N. Popov, and V. Congradac, “Soft sensor for real-time cement fineness estimation,” *ISA Trans,* vol. 55, pp. 250-9, Mar, 2015.

[8] L. Weitao, W. Dianhui, and C. Tianyou, “Multisource Data Ensemble Modeling for Clinker Free Lime Content Estimate in Rotary Kiln Sintering Processes,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 45, no. 2, pp. 303-314, 2015.

[9] M. Sadeghian, and A. Fatehi, “Identification, prediction and detection of the process fault in a cement rotary kiln by locally linear neuro-fuzzy technique,” *Journal of Process Control,* vol. 21, no. 2, pp. 302-308, 2011.

[10] J. Qiao, and T. Chai, “Soft measurement model and its application in raw meal calcination process,” *Journal of Process Control,* vol. 22, no. 1, pp. 344-351, 2012.

[11] J.-s. Wang, N.-n. Shen, X.-d. Ren, and G.-n. Liu, “RBF Neural Network Soft-Sensor Modeling of Rotary Kiln Pellet Quality Indices Optimized by Biogeography-Based Optimization Algorithm,” *Journal of Chemical Engineering of Japan,* vol. 48, no. 1, pp. 7-15, 2015.

[12] S. Lu, H. Yu, H. Dong, X. Wang, and Y. Sun, “Single-step prediction method of burning zone temperature based on real-time wavelet filtering and KELM,” *Engineering Applications of Artificial Intelligence,* vol. 70, pp. 142-148, 2018.

[13] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological cybernetics,* vol. 36, pp. 193-202, 02/01, 1980.

[14] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, “Object Classification Using CNN-Based Fusion of Vision and LIDAR in Autonomous Vehicle Environment,” *IEEE Transactions on Industrial Informatics,* vol. 14, no. 9, pp. 4224-4231, 2018.

[15] K. Wang, K. Li, L. Zhou, Y. Hu, Z. Cheng, J. Liu, and C. Chen, “Multiple convolutional neural networks for multivariate time series prediction,” *Neurocomputing,* vol. 360, pp. 107-119, 2019.

[16] J. Hopfield, “Neural network and physical systems with collective computational abilities,” *Proceedings of The National Academy of Sciences - PNAS,* vol. 79, 01/01, 1982.

[17] Y. Cheng, H. Zhu, J. Wu, and X. Shao, “Machine Health Monitoring Using Adaptive Kernel Spectral Clustering and Deep Long Short-Term Memory Recurrent Neural Networks,” *IEEE Transactions on Industrial Informatics,* vol. 15, no. 2, pp. 987-997, 2019.

[18] S. Hochreiter, and J. Schmidhuber, “Long Short-term Memory,” *Neural computation,* vol. 9, pp. 1735-80, 12/01, 1997.

[19] S. Hochreiter, “The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, pp. 107-116, 04/01, 1998.

[20] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” 06/03, 2014.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” 12/11, 2014.

[22] Z. Kong, B. Tang, L. Deng, W. Liu, and Y. Han, “Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units,” *Renewable Energy,* vol. 146, pp. 760-768, 2020.

[23] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, “Cloud-Assisted Multi-View Video Summarization using CNN and Bi-Directional LSTM,” *IEEE Transactions on Industrial Informatics*, pp. 1-1, 2019.

[24] T.-Y. Kim, and S.-B. Cho, “Predicting residential energy consumption using CNN-LSTM neural networks,” *Energy,* vol. 182, pp. 72-81, 2019.

[25] X. Wang, H. Chen, U. Sultan, X. Zhu, Z. Wang, and G. Xiao, “A Brief Review of the Combustion Diagnosing Techniques for Coal-Fired Boilers of Power Plants in China,” *IEEE Access,* vol. 7, pp. 126127-126136, 2019.

[26] P. Zhou, D. Guo, H. Wang, and T. Chai, “Data-Driven Robust M-LS-SVR-Based NARX Modeling for Estimation and Control of Molten Iron Quality Indices in Blast Furnace Ironmaking,” *IEEE Trans Neural Netw Learn Syst,* vol. 29, no. 9, pp. 4007-4021, Sep, 2018.

[27] Y. Lecun, “PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models),” *Universite P. et M,* vol. Curie (Paris 6), 1987.

[28] K. M. Hornik, M. Stinchcomb, and H. White, “Multilayer feedforward networks are universal approximator,” *IEEE Transactions on Neural Networks,* vol. 2, 01/01, 1989.

[29] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, “Phoneme recognition using time-delay neural networks,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 37, pp. 328-339, 04/01, 1989.

[30] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, “Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks,” *IEEE Transactions on Industrial Electronics,* vol. 65, no. 2, pp. 1539-1548, 2018.